Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 207–214

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE)

J.C. Bailón-Elvira[a], M.J. Cobo[b], E. Herrera-Viedma[a], A.G. López-Herrera[a,1]

[a]*Dept. of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain*
[b]*Dept. Computer Science and Engineering, University of Cádiz, Avenida Ramón Puyol, 11202, Algeciras, Cádiz, Spain.*

Abstract

Since Internet was born most people can access fully free to a lot sources of information. Every day a lot of web pages are created and new content is uploaded and shared. Never in the history the humans has been more informed but also uninformed due the huge amount of information that can be access. When we are looking for something in any search engine the results are too many for reading and filtering one by one. Recommended Systems (RS) was created to help us to discriminate and filter these information according to ours preferences.

This contribution analyses the RS of the official agency of publications in Spain (BOE), which is known as *"Mi BOE"*. The way this RS works was analysed, and all the meta-data of the published documents were analysed in order to know the coverage of the system. The results of our analysis show that more than 89% of the documents cannot be recommended, because they are not well described at the documentary level, some of their key meta-data are empty. So, this contribution proposes a method to label documents automatically based on Latent Dirichlet Allocation (LDA). The results are that using this approach the system could recommend (at a theoretical point of view) more than twice of documents that it now does, 11% vs 23% after applied this approach.

Keywords: Recommender systems, BOE, LDA, Alerts

1. Introduction

Nowadays there exist a huge amount of information that can be accessed trough Internet but we can not read all results returned by any search engine, neither filter one by one. Recommender Systems (RS) can perform this task of filtering for us. There are a lot of models and approaches about RS, but in essence what they want is know the users preferences and offer a list of relevant items bases on these preferences.

RS can be divided in three main groups according how they filter the information:

*Corresponding author. Tel.: +34-958-248557; fax: +34-958-243317.
E-mail address: lopez-herrera@decsai.ugr.es.

1877-0509
10.1016/j.procs.2019.11.277

- Content Based: This approach gets the preferences of the user and search similar items.
- Collaborative Filtering: the system gets the rank of the each item provided by the user and predict the utility of these items for others similar users.
- Hybrid Systems: these systems use techniques based on both Content and Collaborative Filtering. The objective of this type of systems is to combine the benefits of both approaches

In the literature can be found a lot of examples about RS. In medicine and health care exist different RS [1, 2, 3, 4], others RS help to search which holidays places could like us [5, 6], which music might like us [7], places to visit while we are travelling [8], films or series to watch [9, 10, 11, 12], news [13], on line market [14, 15], multimedia resources on social networks [16], possible team mates to work [17], papers to read [18, 19] or methodologies to apply into a class [20, 21, 22].

The official agency of publications in Spain (BOE) publishes the documents generated by the Spanish government since 1960 until today, every day of the week from Monday to Saturday. Nowadays millions of documents has been published and can be freely accessed by the citizens. This huge amount of documents can not been filtered by a human and here is where RS are needed. The agency has its own RS, which is called *"Mi BOE"*. The users declare about what are they interested for in the tabs of the preference page (Fig. 1), the system will send you a list of documents (as they are published), according to your selected preferences. The idea is that the user does not have to search daily for their documents of interest, the user delegates this task in the RS "Mi BOE".

Fig. 1. *"Mi BOE"* configuration page.

The words showed in Fig. 1 belong to a portion of two lists used to describe the published documents. It is on the basis of these words that the system works. At the time of recommending the system searches the words selected by the user in the configuration page and matches them with the words that appear in the meta-data *alertas* (alerts) and *materias* (subject-matters) of the documents in question. The problem is that many (too many) documents do not have values in this meta-data, i.e. their subject-matters and alerts are empty, as shown in the following scenario. Lets suppose a student which is waiting to apply for a researcher position at the university. That student logs into "Mi BOE" and selects *Becas* (Grants) as subject-matter. Then he/she waits achieve a list with documents about research grants. The problem becomes when some of the documents related to grants never be recommended because they are not described (meta-data is empty). The Listing 1 is a real example of empty documents, which the system never will list to the student. If the student fully would rely in this RS, he/she never will apply to this grant and lost this research opportunity. For comparison, Listing 2 shows a non-empty document, which presents values for three subject-matters (*'Abogados'*, *'Planes de estudios'* and *'Universidad Autónoma de Madrid'*).

```
-------------------------------------------------------------------------------
| <identificador>BOE-B-2018-49522</identificador>
| <titulo>Extracto de la Resoluci\'on de 15 de octubre de 2018, de la Secretaría
| de Estado de Universidades, Investigaci\'on, Desarrollo e Innovaci\'on por
| la que se convocan ayudas complementarias para el año 2019, destinadas a
| beneficiarios del subprograma de formación del profesorado universitario.</titulo>
| ...
| <materias/>
| <materias_cpv/>
-------------------------------------------------------------------------------
```

Listing 1. Extract of XML of the empty document BOE-B-2018-49522 about university grants.

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2017-9252</identificador>
| <titulo>Resolución de 14 de julio de 2017, de la Universidad Autónoma de Madrid,
| por la que se publica la modificación del plan de estudios de Máster en Acceso
| a la Profesión de Abogado.</titulo>
| ...
| <materias>
| <materia codigo="8" orden="">Abogados</materia>
| <materia codigo="5605" orden="">Planes de estudios</materia>
| <materia codigo="7016" orden="">Universidad Autónoma de Madrid</materia>
| </materias>
| <alertas/>
-------------------------------------------------------------------------------
```

Listing 2. Extract of XML of the non-empty document BOE-A-2017-9252 about master curriculum.

It is very important that the documents are well described, but we can see in the agency's information system there are too many empty documents. These documents have to be described, more than a million have neither alerts nor subject-matters and this is a huge problem. So, this contribution proposes an automatic method for filling up the empty meta-data, which is based on the application of Latent Dirichlet Allocation (LDA).

The rest of this work is organised as follows. Section *Methodology* summaries the main steps carried out. In Section *Analysis* the study of the meta-data is shown. Section *Model and Evaluation* presents the experimentation carried out the main results. Finally, some conclusions are drawn.

2. Methodology

The steps carried out in this contribution were:

1. **Analyse the content of the BOE's information system**: it was studied the way to download the documents published by the agency from 1960 to 2018 (both includes). For this task a crawler was created, which reads each HTML page published and scraped the XML format. Next the documents and the meta-data was saved into a local relational database.
2. **Data analysis**: the saved meta-data was analysed to know the different key elements that the BOE uses for describing their documents.
3. **Automatic topic modelling**: using LDA to self labelling the documents not described.
4. **Evaluate the model**: in order to know how the application of LDA improves the current RS *"Mi BOE"*.

3. Analysis

The BOE publishes every day (from Monday to Saturday) the documents approved by the different governmental agencies in Spain. These documents are related about laws, orders, announcements, grants, etc. Among the different document types offered for each document (PDF, ePub and XML), the XML version offers a structured format of the content (meta-data), which is easily processed by computers. Each XML has a defined fields that can be filled or not. Some of these fields are: the ID of the document, published date, full text, title, references

about other documents, etc.

Among all analysed meta-data exist two fields which are used to describe the content of the documents. These meta-data are *alertas* (alerts) and *materias* (subject-matters). A document can have 0 *Alertas* or uses still 6 words to describe the documents. On the other hand, up to 25 words as subject-matters may be used to describe a document or none at all. 42 different alerts can be used, and more than 6000 different terms as subject-matters.

The amount of documents published from 1960 to 2018 (both inclusive) are around 1.400.000. The documents described by *alertas* are 82.923, and those described by *materias* are 132.042. The Fig. 2 shows the percentage of documents described and not described. More than 89% of documents are not described (they are empty) and nearly 11% are described.

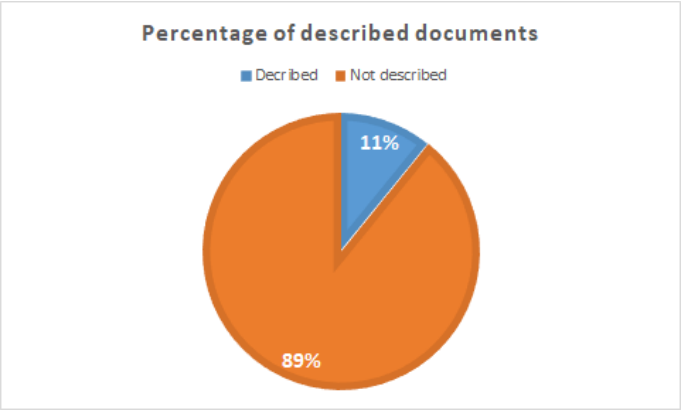

Fig. 2. Percentage of described and not described documents.

The analysis shows a high percentage of documents without descriptors and this is a big problem for the RS used by BOE, because the documents without *materias* neither *alertas* never could be recommended. In the section 4 an approach to partially solve this problem is proposed, it is based on the use of the LDA technique.

4. Model and Evaluation

In order to improve the RS *"Mi BOE"* the use of Latent Dirichlet Allocation (LDA) [23] is carried out. In this proposal the R language [24] is used together the RWeka package [25]. LDA is an algorithm which can detects the topics into the documents analysing their content. As example, having a training set of documents about fuzzy logic the algorithms return a list of relevant words about the topic 'fuzzy logic'. Using these relevant words knows as bag-word the algorithm can detect if a new document could be about fuzzy logic or not.

In this work the meta-data selected as information source was *titulo* (title), where it is a brief abstract about the document, together the list of 42 unique *alertas*. Using these two meta-data was created the data training with 79.409 documents. For each *alerta* was retrieved a list of documents that they had it and was processed as follows (see top Fig. 3):

1. Extract the meta-data *titulo* from all documents.
2. Clean the meta-data *titulo* removing dots, stop-words, numbers, extra spaces, and setting all text to lowercase. Next, tokenize using n-grams with minimum length of 1 and 3 as maximum.
3. Create and save the bag-word associated for each *alerta*.

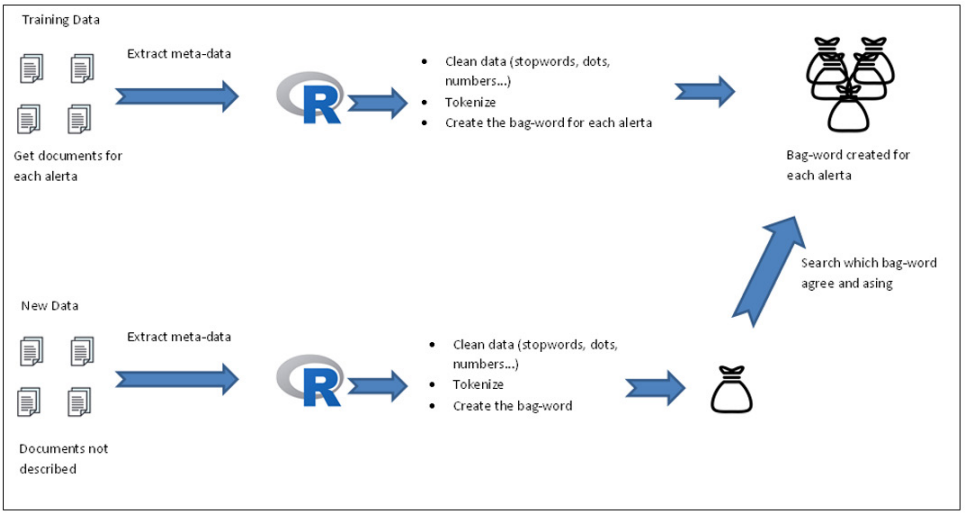

Fig. 3. LDA process. At the top the process to *alertas*. At bottom the process to the described documents.

Finished the process of creating the bag-word for each *alerta* the next step was get the 1.299.409 empty documents and to apply the steps show at bottom of Fig. 3. The steps are:

1. Extract the meta-data *titulo* from the document.
2. Clean the meta-data *titulo* removing dots, stop-words, numbers, extra spaces and finally set all text to lowercase. Next, tokenize using n-grams with minimum length of 1 and 3 as maximum.
3. Create the bag-word of the document.
4. Compare the bag-word created against the bag-words of the *alertas*. If it matches then that *alerta* is assigned to the document.

As example of working, the Listing 3 shows the extract of the document with identification BOE-A-2017-9230[1] where its descriptors (*materias* and *alertas*) are empty. According the title of this document its content is about the designation of a person as a Consumer Council. The BOE system has an alert (*alerta*) related with this topic, which is called *'Nombramientos y ceses de altos cargos'* ('Appointments and resignations of senior officials'). After the LDA process ends, this documents is enriched with the correct alert.

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2017-9230</identificador>
| <titulo>Orden SSI/750/2017, de 19 de junio, por la que se nombra vocal del Consejo
| de Consumidores y Usuarios a don José Ángel Oliván García.</titulo>
| ...
| <materias/>
| <alertas/>
-------------------------------------------------------------------------------
```

Listing 3. Extract of XML of empty document BOE-A-2017-9230.

Other example of empty document is the one which identification is BOE-A-2011-6485[2]. The Listing 4 shows the no key meta-data for this document. Reading the content of this document, it treats about a public competitive process in a university position. For this kind of documents there exists the *alerta 'Oposición'*, with the bag-words composed by: *'plaz', 'univers', 'nombr', 'convocatori', 'referent'*. Using LDA with this document the model back the next list: *'acces', 'docent', 'universitari'*. As the before example the system detected a match and assigned the correct *alerta* to this document.

[1] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2017-9230
[2] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2011-6485

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2011-6485</identificador>
| <titulo>Resolución de 23 de marzo de 2011, de la Universidad de Zaragoza, por
| la que se corrigen errores en la de 4 de marzo de 2011, por la que se convoca
| concurso de acceso a plaza de cuerpos docentes universitarios.</titulo>
| ...
| <materias/>
| <alertas/>
-------------------------------------------------------------------------------
```

Listing 4. Extract of XML of document BOE-A-2011-6485.

If this LDA based method is applied against the not described documents in the BOE corpus, the amount of described documents increases in 193.589. This represents the 13% of the entire collection of the BOE and it increases twice the documents that actually can be recommend by the RS *"Mi BOE"*. Fig. 4 shows the percentage of documents described after LDA proccess. The RS could recommend nearly 25% of the entire collection.

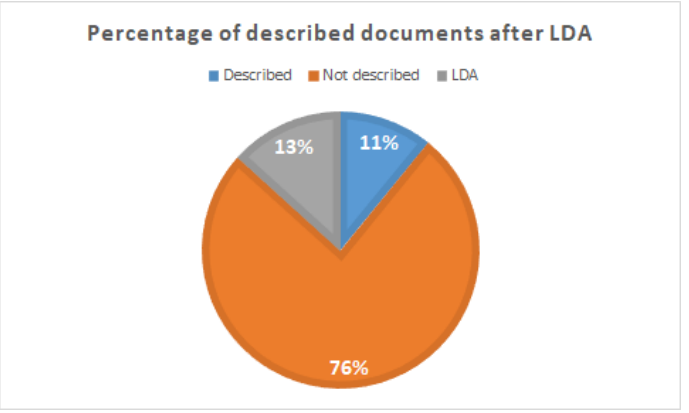

Fig. 4. Documents described after LDA process.

Using the LDA model the described documents were evaluated in order to know if the model could label these documents by right way. The result was hopeful and shows that the published documents in the last years were described more accurate (the topics automatically assigned to previously labelled documents concordat in a 69%).

As well the results show that in some documents the LDA model assigned more descriptors than the originals had, for example, the original document BOE-A-2015-76 [3] (see Listing 5) is about job and teaching in universities had only an alert *Oposiciones* ('Compettitive examination') and the model assigned two, *Oposiciones* ('Competitive examination') and *Educación y enseñanza* ('Education and teaching').

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2015-76</identificador>
| <titulo>Resolución de 25 de noviembre de 2014, conjunta de la |Universidad de
|Granada y del Servicio Andaluz de Salud, por la que se convoca
|concurso de acceso a plazas vinculadas de cuerpos docentes |universitarios.</titulo>
| ...
| <materias/>
| <alertas>
| <alerta codigo="140"orden="">Oposiciones</alerta>
| </alertas>
-------------------------------------------------------------------------------
```

Listing 5. Extract of XML of document BOE-A-2015-76.

[3] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2015-76

In other cases the system assigns terms such as *Oposiciones* ('Competitive examination') instead of *Concursos de personal público* ('Recruitment competitions to the State'). Both terms are semantically very similar. This is the case of document BOE-A-2015-73[4] (see Listing 6).

```
-------------------------------------------------------------------------------
| <identificador>BOE-A-2015-73</identificador>
| <titulo>Resolución de 17 de diciembre de 2014, del Ayuntamiento de Cambre
| (A Coruña),referente a la convocatoria para proveer puesto de trabajo por el
|sistema de concurso.</titulo>
| ...
| <materias/>
| <alertas>
|<alerta codigo="141" orden="">Concursos de personal público</alerta>
|</alertas>
-------------------------------------------------------------------------------
```

Listing 6. Extract of XML of document BOE-A-2015-73.

5. Conclusions

RS are needed in order to help the users to filter the information returned by the conventional information retrieval systems. In this work an approach for improving the RS called *"Mi BOE"*, developed by the official agency of publications in Spain (BOE), was proposed.

"Mi BOE" recommends documents according to the explicit preferences selected by the users, the system shows a list of documents that matches which the user's preferences. These preferences are the topics of the documents, which are called *materias* (subject-matters) and *alertas* (alerts) in the BOE's terminology. More than 89% of the documents are not described (they are empty), so the RS can not recommend a lot of documents. *"Mi BOE"* could recommend up to 11% of the entire collection, a percentage very very low.

In this work the use of Latent Dirichlet Allocation (LDA) method to assign automatically descriptors based on *alertas* is proposed. This approach shows how the system could recommend about 25% of documents of the entire collection, improving more than twice the actual performance of this RS. In this way this approach shows how LDA can improve the task of automatically describing empty documents.

The evaluation shows that the proposed model has good results and enhanced the documentary description.

As future works our desire is to solve the problems with the descriptors that are semantically similar, for example, using an ontology and normalising these terms. Also creating a web interface in order to use real users and evaluate the proposal in a more realistic environment.

Also we plan increase more than the 25% labelling achieved in this proposal and we will try to describe the complete collection with automatic methods.

References

[1] M. Hung, J. Xu, E. Lauren, M. W. Voss, M. N. Rosales, W. Su, B. Ruiz-Negron, Y. He, W. Li, F. W. Licari, Development of a recommender system for dental care using machine learning, SN APPLIED SCIENCES 1 (7). doi:{10.1007/s42452-019-0795-7}.

[2] B. Esteban, Á. Tejeda-Lorente, C. Porcel, M. Arroyo, E. Herrera-Viedma, TPLUFIB-WEB: A fuzzy linguistic Web system to help in the treatment of low back pain problems, Knowledge-Based Systems 67 (2014) 429–438. doi:10.1016/j.knosys.2014.03.004.

[3] C. Suphavilai, D. Bertrand, N. Nagarajan, Predicting Cancer Drug Response using a Recommender System, BIOINFORMATICS 34 (22) (2018) 3907–3914. doi:{10.1093/bioinformatics/bty452}.

[4] J. M. Morales-del Castillo, E. Peis, A. A. Ruiz, E. Herrera-Viedma, Recommending biomedical resources: A fuzzy linguistic approach based on semantic web, International Journal of Intelligent Systems 25 (12) (2010) 1143–1157.

[4] https://www.boe.es/diario_boe/xml.php?id=BOE-A-2015-73

[5] J. P. Lucas, N. Luz, M. N. Moreno, R. Anacleto, A. Almeida Figueiredo, C. Martins, A hybrid recommendation approach for a tourism system, Expert Systems with Applications 40 (9) (2013) 3532–3550. doi:10.1016/j.eswa.2012.12.061.
[6] K. McCarthy, K. McCarthy, M. Salamo, M. Salamo, L. Coyle, L. Coyle, L. McGinty, L. McGinty, B. Smyth, B. Smyth, P. Nixon, P. Nixon, CATS: A Synchronous Approach to Collaborative Group Recommendation, Flairs (2006) 86–91.
[7] A. Crossen, J. Budzik, K. J. Hammond, Flytrap: Intelligent group music recommendation, in: Proceedings of the 7th International Conference on Intelligent User Interfaces, IUI '02, ACM, New York, NY, USA, 2002, pp. 184–185. doi:10.1145/502716.502748.
[8] X. Zheng, Y. Luo, L. Sun, J. Zhang, F. Chen, A tourism destination recommender system using users' sentiment and temporal dynamics, JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 51 (3) (2018) 557–578. doi:{10.1007/s10844-018-0496-5}.
[9] S. Amer-yahia, S. B. Roy, A. Chawla, G. Das, C. Yu, Group Recommendation : Semantics and Efficiency, Proceedings of the VLDB Endowment 2 (1) (2009) 754–765. doi:10.1.1.151.4805.
[10] M O'connor, D. Cosley, J. Konstan, J. Riedl, PolyLens: A recommender system for groups of users, Ecscw 2001 (In Proceedings of the European Conference on Computer-Supported Cooperative Work) (2001) 199–218. doi:10.1145/312129.312230.
[11] C. A. Gomez-Uribe, N. Hunt, The Netflix Recommender System, ACM Transactions on Management Information Systems 6 (4) (2015) 1–19. doi:10.1145/2843948.
[12] J. Masthoff, Group modeling: Selecting a sequence of television items to suit a group of viewers, User Modeling and User-Adapted Interaction 14 (1) (2004) 37–85. doi:10.1023/B:USER.0000010138.79319.fd.
[13] H. Wen, L. Fang, L. Guan, A hybrid approach for personalized recommendation of news on the web, Expert Systems with Applications 39 (5) (2012) 5806 – 5814. doi:https://doi.org/10.1016/j.eswa.2011.11.087.
URL http://www.sciencedirect.com/science/article/pii/S0957417411016332
[14] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Computing 7 (1) (2003) 76–80. arXiv:69, doi:10.1109/MIC.2003.1167344.
[15] D. R. Liu, C. H. Lai, W. J. Lee, A hybrid of sequential rules and collaborative filtering for product recommendation, Information Sciences 179 (20) (2009) 3505–3519. doi:10.1016/j.ins.2009.06.004.
[16] F. Amato, V. Moscato, A. Picariello, F. Piccialli, SOS: A multimedia recommender System for Online Social networks, FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE 93 (2019) 914–923. doi:{10.1016/j.future.2017.04.028}.
[17] C. Porcel, A. G. López-Herrera, E. Herrera-Viedma, A recommender system for research resources based on fuzzy linguistic modeling, Expert Systems with Applications 36 (3 PART 1) (2009) 5173–5183. doi:10.1016/j.eswa.2008.06.038.
URL http://dx.doi.org/10.1016/j.eswa.2008.06.038
[18] A. Tejeda-Lorente, C. Porcel, J. Bernab??-Moreno, E. Herrera-Viedma, REFORE: A recommender system for researchers based on bibliometrics, Applied Soft Computing Journal 30 (2015) 778–791. doi:10.1016/j.asoc.2015.02.024.
[19] D. Wang, Y. Liang, D. Xu, X. Feng, R. Guan, A content-based recommender system for computer science publications, KNOWLEDGE-BASED SYSTEMS 157 (2018) 1–9. doi:{10.1016/j.knosys.2018.05.001}.
[20] H. El Fazazi, M. Qbadou, I. Salhi, K. Mansouri, Personalized recommender system for e-Learning environment based on student's preferences, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY 18 (10) (2018) 173–178.
[21] H. Lin, S. Xie, Z. Xiao, X. Deng, H. Yue, K. Cai, Adaptive Recommender System for an Intelligent Classroom Teaching Model, INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN LEARNING 14 (5) (2019) 51–63. doi:{10.3991/ijet.v14i05.10251}.
[22] C. Cobos, O. Rodriguez, J. Rivera, J. Betancourt, M. Mendoza, E. Leon, E. Herrera-Viedma, A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes, Information Processing & Management 49 (2013) 607–625. doi:10.1016/j.ipm.2012.12.002.
[23] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, JOURNAL OF MACHINE LEARNING RESEARCH 3 (4-5) (2003) 993–1022, 18th International Conference on Machine Learning, WILLIAMSTOWN, MASSACHUSETTS, JUN 28-JUL 01, 2001. doi:{10.1162/jmlr.2003.3.4-5.993}.
[24] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2008).
URL http://www.R-project.org
[25] K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R meets Weka, Computational Statistics 24 (2) (2009) 225–232. doi:10.1007/s00180-008-0119-7.